The War Between Mice and Elephants

Liang Guo and Ibrahim Matta
Computer Science Department
Boston University

9th IEEE International Conference on Network Protocols (ICNP), Riverside, CA, November 2001.

Acknowledgements

Figures in this presentation are taken from a class presentation by Matt Hartling and Sumit Kumbhar in CS577 Advanced Computer Networks in Spring 2002.

Outline

- Introduction and Motivation
- Performance Metrics
- Active Queue Management
  - Drop Tail, RED and RIO Routers
  - *DiffServ: Core versus Edge Routers*
- Proposed Architecture
- Analysis via ns-2 simulation
- Discussion
- Conclusions

Introduction

- 80% of the traffic is due to a small number of flows {elephants} .
- The remaining traffic volume is due to many short-lived flows {mice} .
- With TCP congestion control mechanisms, these short flows receive <u>less</u> than their fair share when they compete for the bottleneck bandwidth.

Introduction

The research goal

- Provide long-lived flows with expected data rate.
- Provide better-than-best-effort service for short TCP flows {Web traffic}.

Introduction

What did the authors do?

- Proposed a new *DiffServ* style architecture designed to be fairer to short flows.
- Ran extensive simulations to demonstrate the value of the proposed scheme.

Performance Metrics

- **Object response time** – *the time to download an object in a Web page.*
- **Transmission time** – the time to transmit a page.
- **goodput** (Mbps) - the rate at which packets arrive at the receiver. *Goodput differs from throughput in that retransmissions are excluded from goodput.*

Performance Metrics

- Jain's fairness
  - For any given set of user throughputs ($x_1, x_2, \ldots, x_n$), the fairness index to the set is defined:

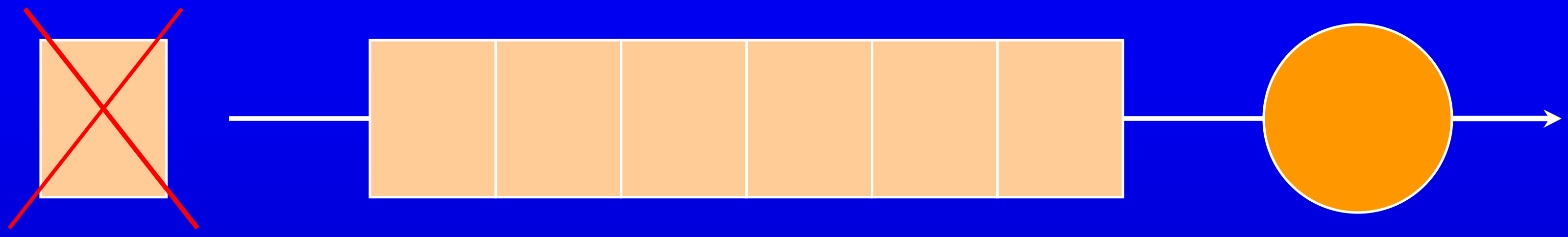
$$f(x_1, x_2, \ldots, x_n) = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2}$$

- Instantaneous queue size – provides a measure of the delay.
- Packet drop/mark rate – rate at which packets are dropped at bottleneck router.

Active Queue Management

- TCP sources interact with routers to deal with congestion caused by an internal bottlenecked link.
- Drop Tail :: FIFO queuing mechanism.
- RED :: Random Early Detection
- RIO :: RED with In and Out

Drop Tail Router

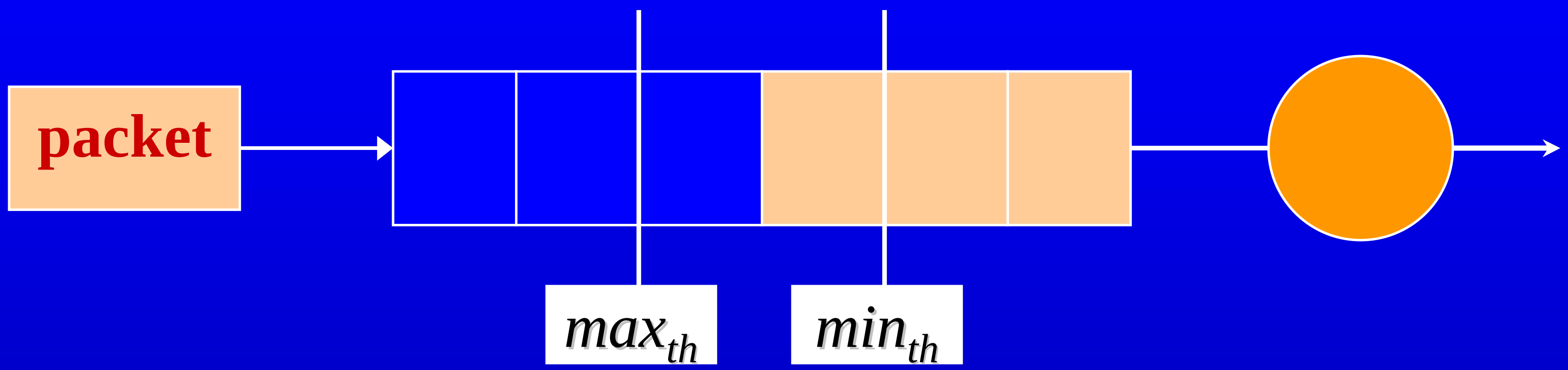


- FIFO queueing mechanism that drops packets when the queue overflows.
- Introduces *global synchronization* when packets are dropped from several connections.

RED Router

- Random Early Detection (RED) detects congestion "early" by maintaining an exponentially-weighted average queue size.
- RED probabilistically drops packets before the queue overflows to signal congestion to TCP sources.
- RED attempts to avoid global synchronization and bursty packet drops.

RED



min_{th} :: average queue length threshold for triggering probabilistic drops/marks.

max_{th} :: average queue length threshold for triggering forced drops.

RED Parameters

q_{avg} :: average queue size

$q_{avg} = (1-w_q) * q_{avg} + w_q *$ instantaneous queue size

w_q :: weighting factor $0.001 <= w_q <= 0.004$

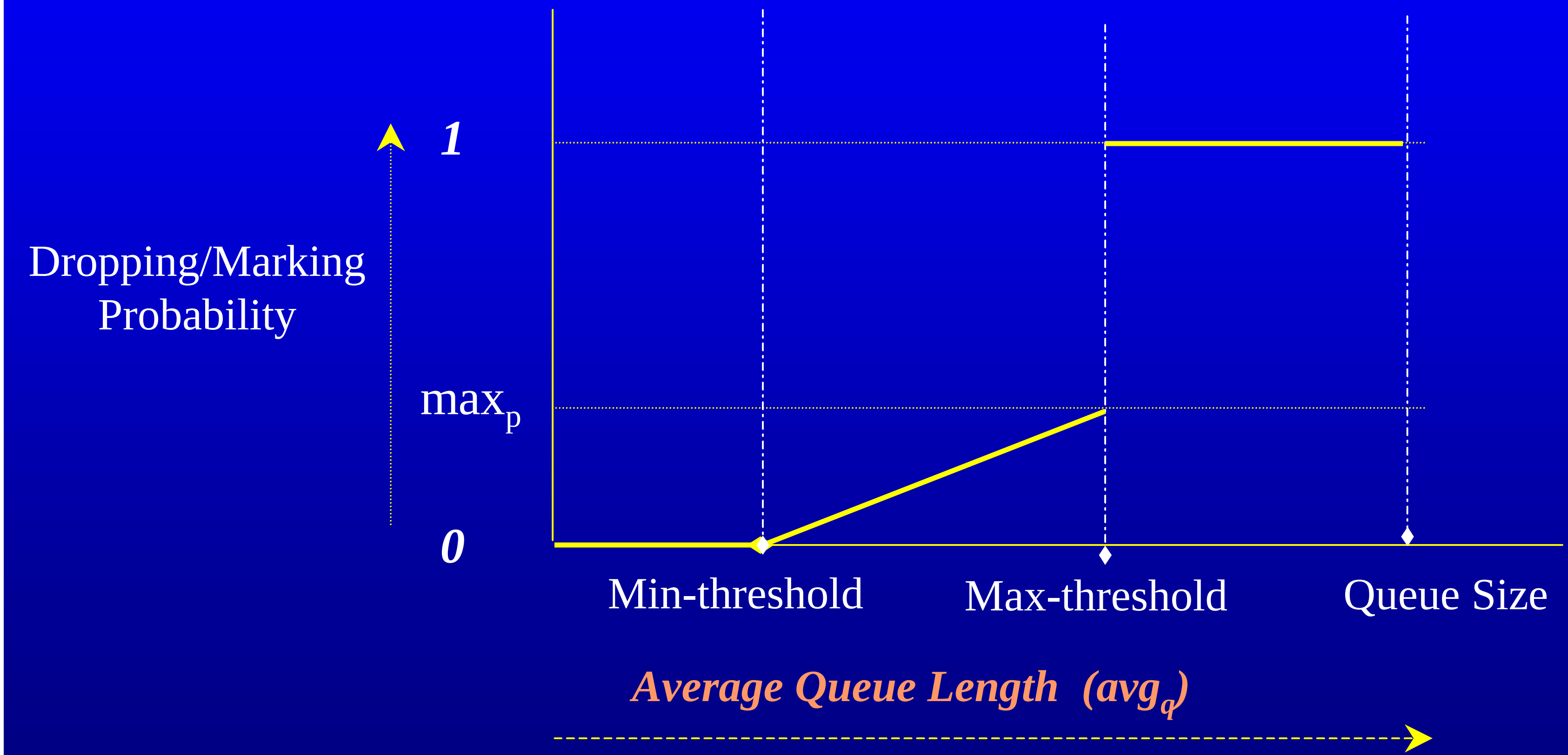
max_p :: maximum dropping/marking probability

$$p_b = max_p * (q_{avg} – min_{th}) / (max_{th} – min_{th})$$

$$p_a = p_b / (1 – count * p_b)$$

buffer_size :: the size of the router queue in packets.

RED Router Mechanism



RIO

- RED with two flow classes (short and long flows)
- There are two separate sets of RED parameters for each flow class.
- Only one real queue exists to avoid packet reordering.
- For long flows, average queue size of total queue is used (Q_{total}).

RIO-PS

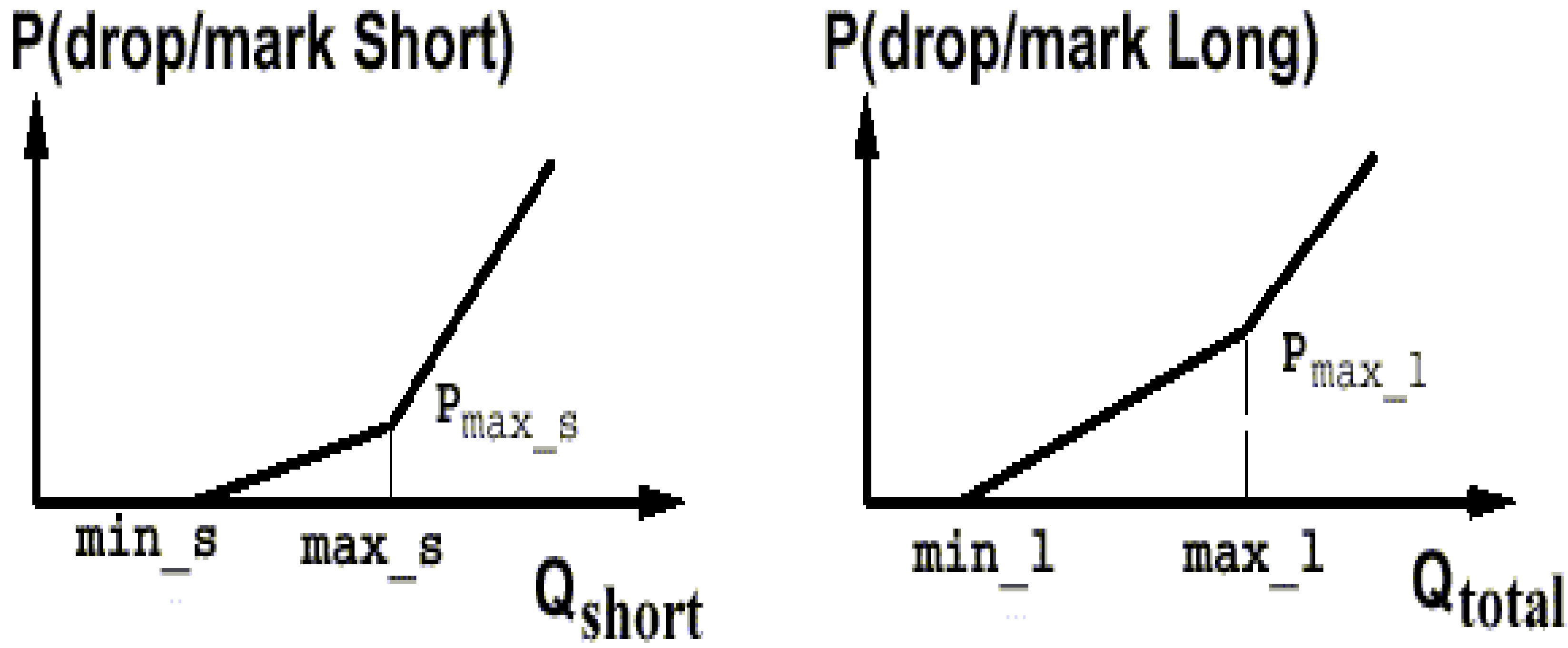


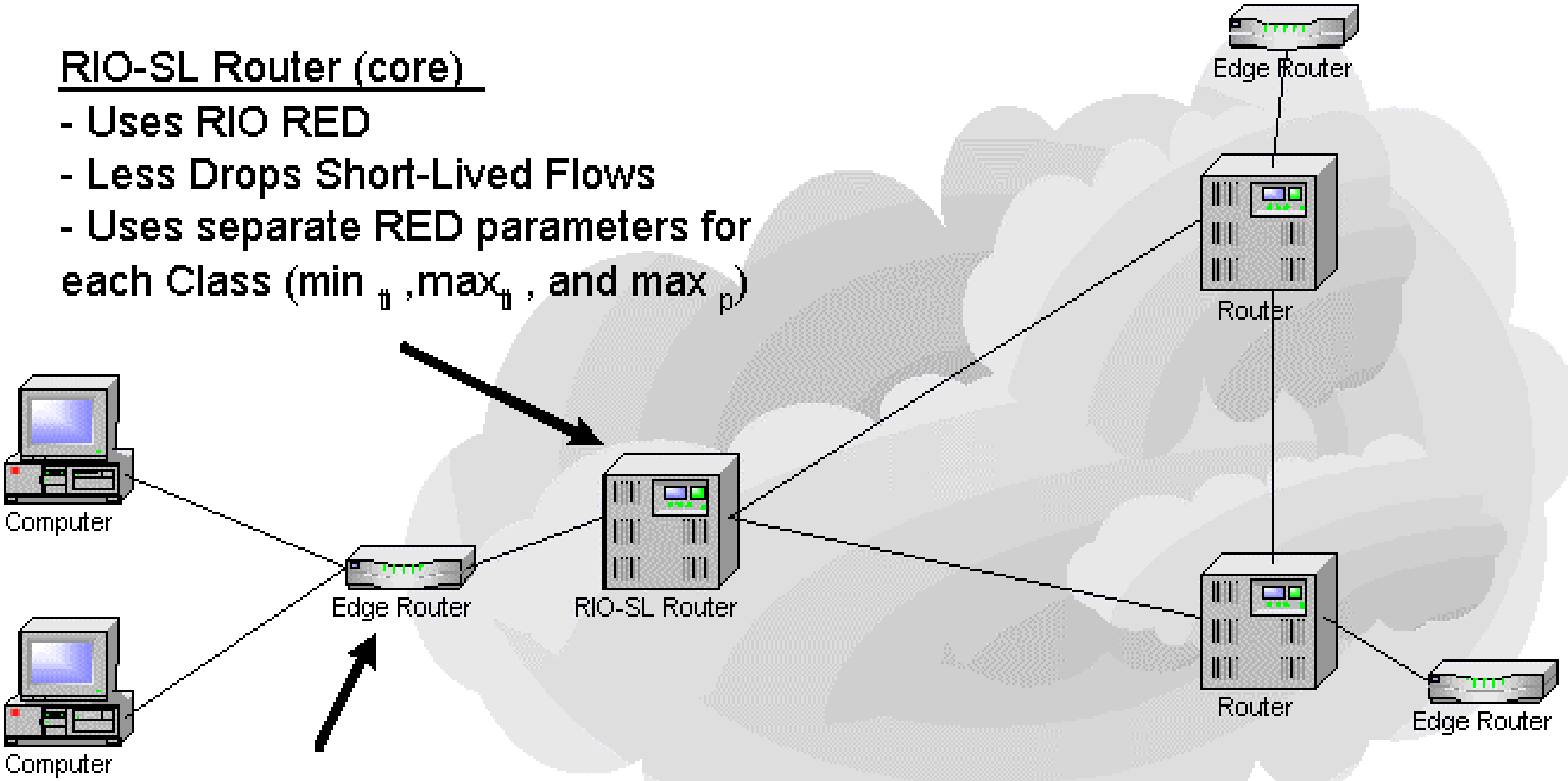
Fig. 4. RIO queue with Preferential treatment to Short flows

DiffServ Philosophy

- Routers divided into edge and core routers.
- Intelligence pushed out to edge (ingress and egress) and core routers are to be "simple".
- Edge router 'classifies' flows and tags packet with classification (e.g., short or long).
- The tag is used by RIO in core router to yield RIO-PS {Preferential treatment for Short flows} .

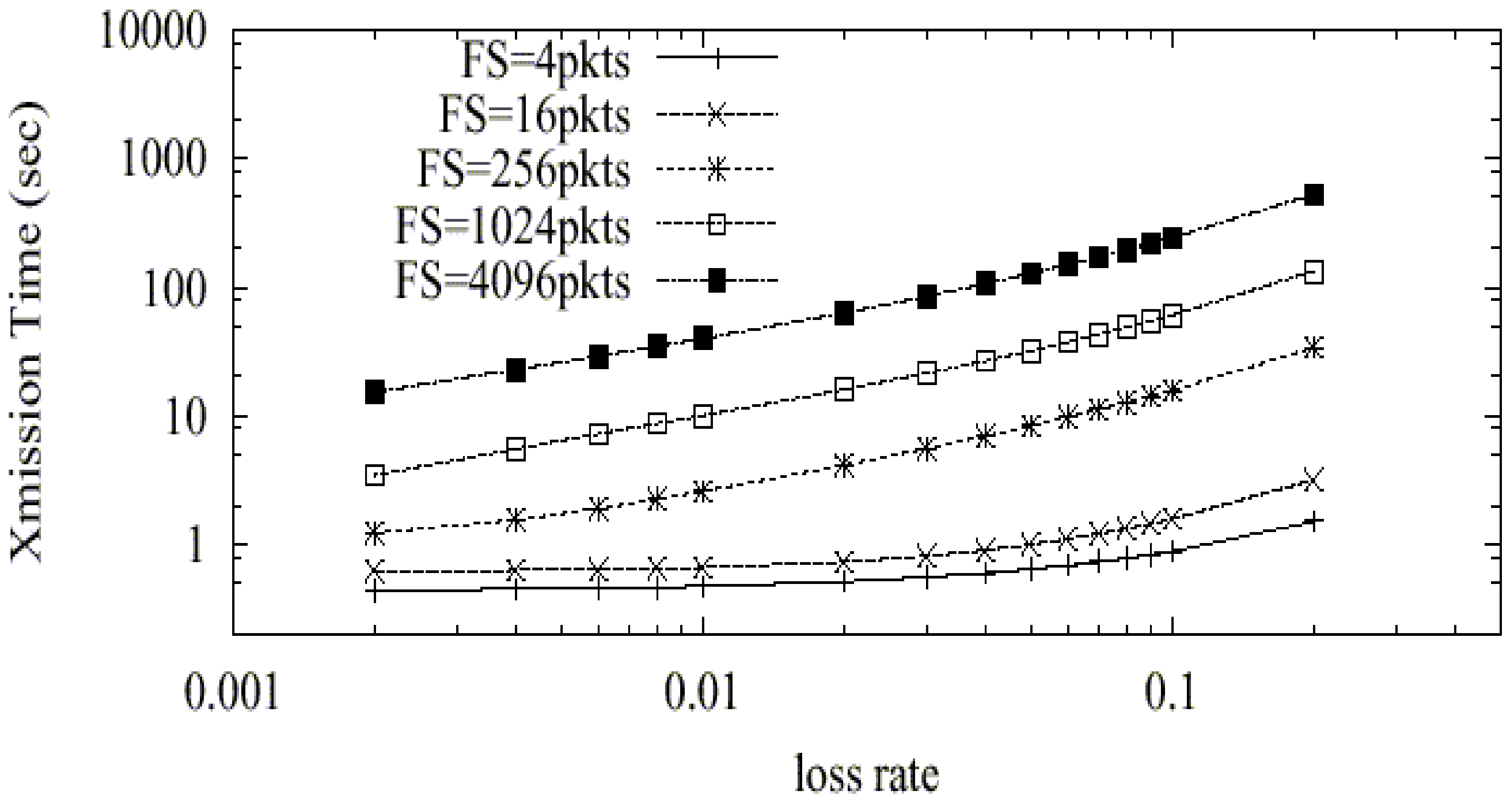
RIO-PS

RIO-SL Router (core)
- Uses RIO RED
- Less Drops Short-Lived Flows
- Uses separate RED parameters for each Class (min_{th}, max_{th}, and max_p)



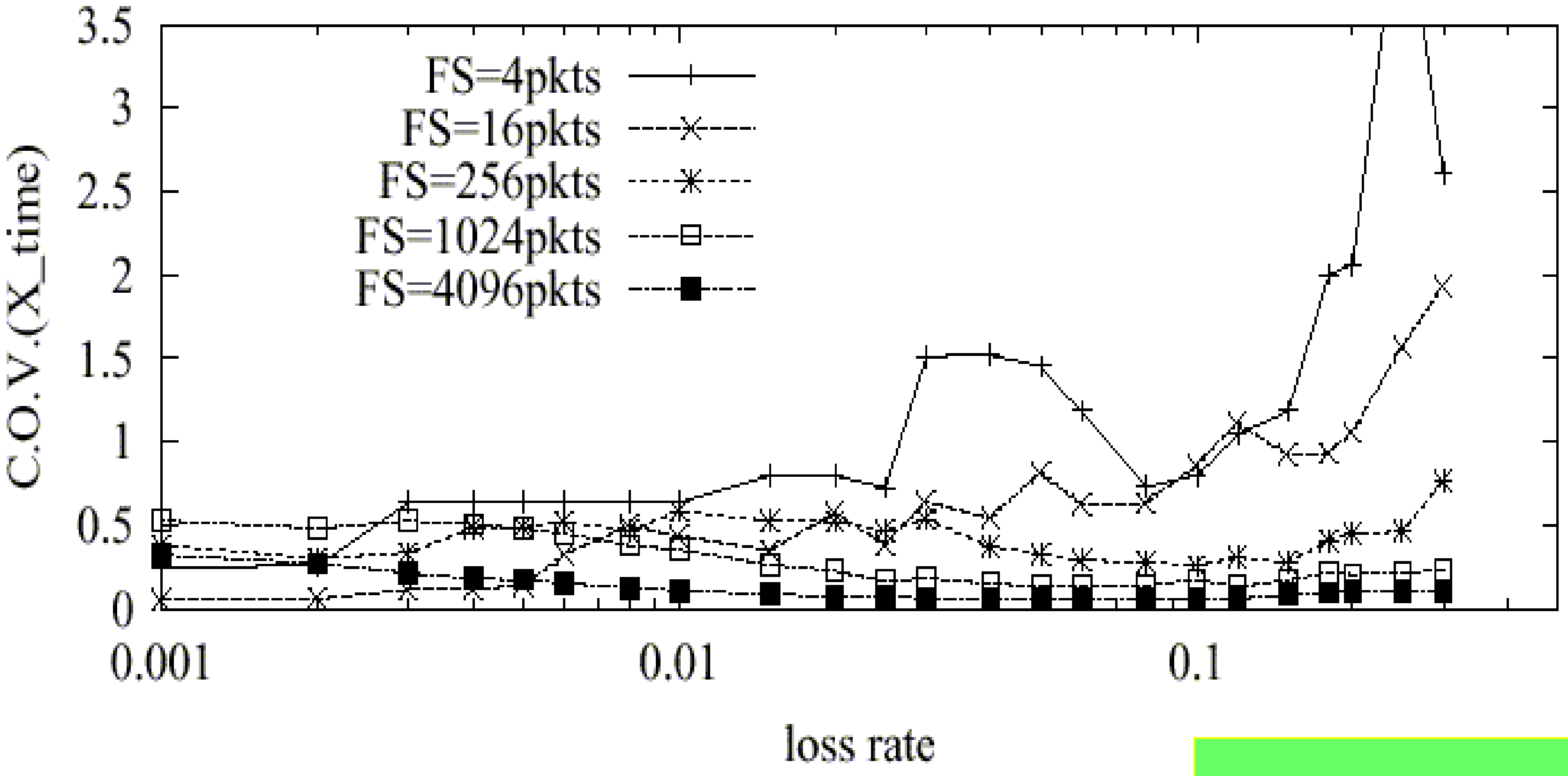
Edge Router
- Classifies Packets as Short or Long
- Perflow State Packet Count

Fig 1a. Average Transmission Time



(a) Average Transmission Time

Fig 1b. Transmission Time Variance



(b) Coefficient of Variation

Conclusion: Reducing the loss probability is more critical to helping the short flows.

Figure 2: Comparison of Drop Tail, RED, RIO-PS

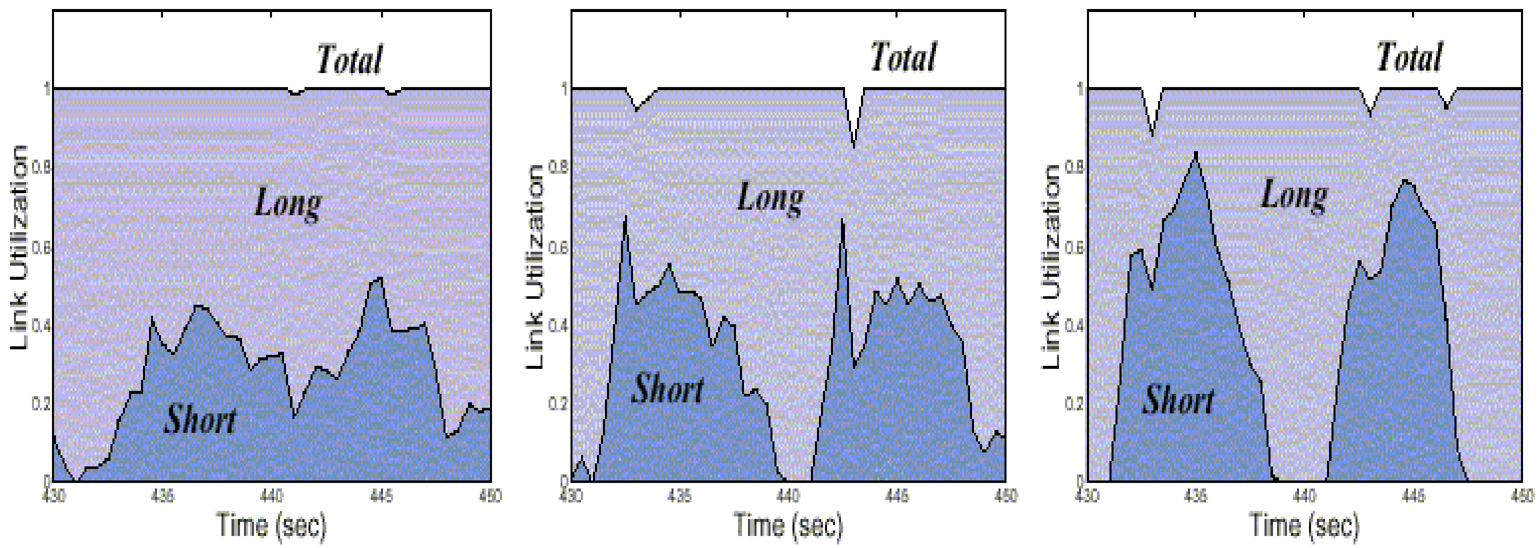


Fig. 2. Impact of Preferential Treatment— Link utilization under Drop Tail (left), RED (middle), and RIO-PS (right)

Table I Goodput

Link B/W	Flows	DropTail	RED	RIO-PS
1.25Mbps	All	153479	154269	154486
	Short	40973	49897	49945
	Long	112506	104372	104541
1.5Mbps	All	185650	184315	183154
	Short	43854	49990	49990
	Long	141796	134325	133164

TABLE I

NETWORK GOODPUT UNDER DIFFERENT SCHEMES

Proposed Architecture

- Edge router classifies flows as belonging to short flow class or long flow class and places tag into packet.
- The edge router uses a threshold L_t and a per flow counter. This per-flow state information is “softly” maintained at the edge router.
- Once the counter exceeds the threshold, the flow is considered a Long flow. The first L_t packets are classified as part of a Short flow.

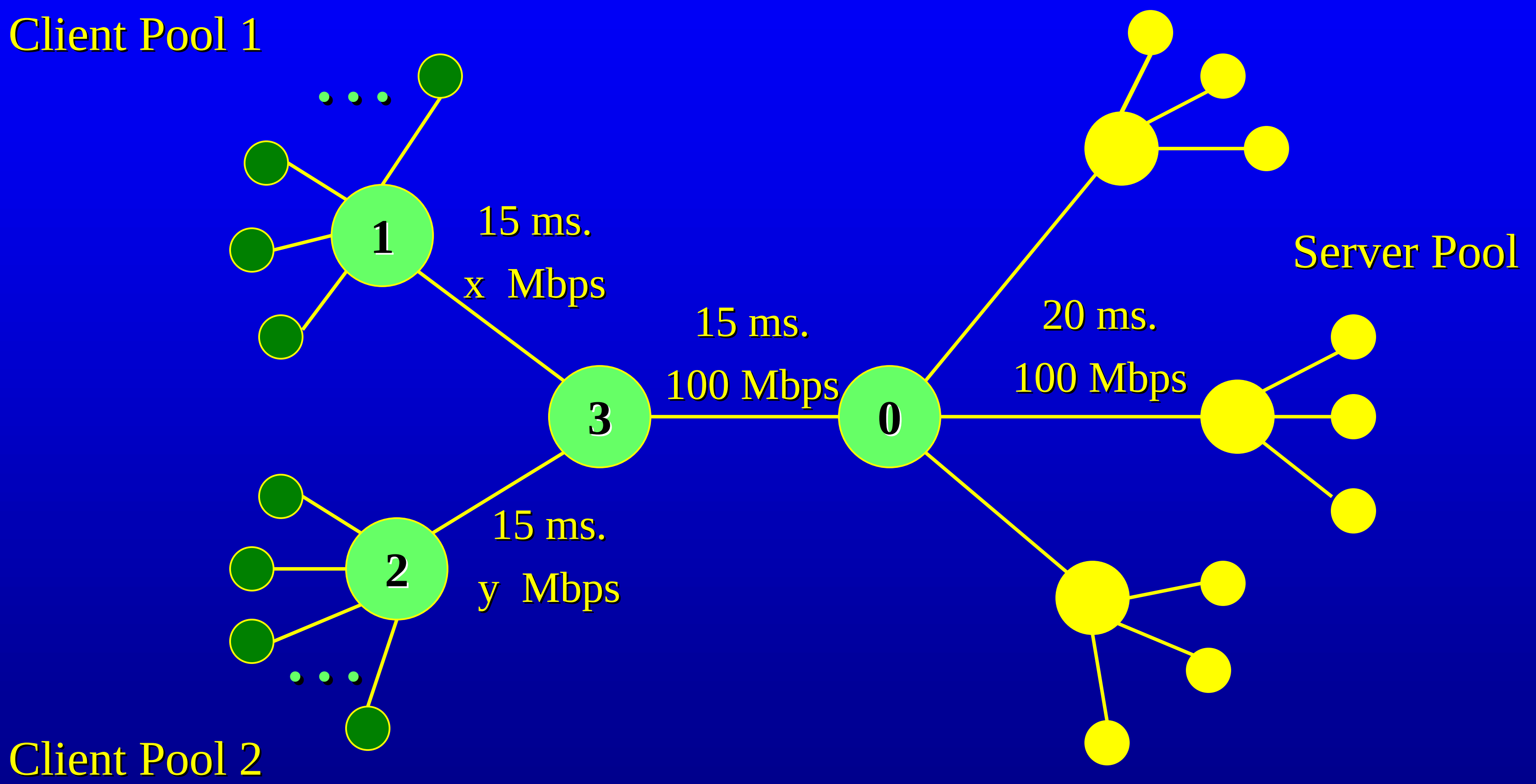
Proposed Architecture

- The threshold can be static or dynamic.
- Dynamic version can be controlled by a desired SLR (Short-to-Long Ratio).
- Core routers give preferential treatment to short flows (e.g. in Table III $p_{max\_s} = 0.05$).

Web Traffic Characterization

- Used Feldman's model in ns-2 simulations:
  - HTTP1.0
  - Exponential inter-page arrivals
  - Exponential inter-object arrivals
  - Uniform distribution of objects per page with min 2 and max 7
  - Object size; bounded Pareto distribution with minimum 4 bytes, maximum 200KB, shape =1.2

Simulation Topology



Description	Value
Packet Size	500 bytes
Maximum Window	128 packets
TCP version	Newreno
TCP timeout Granularity	0.1 seconds
Initial Retransmission Timer	3.0 seconds
B/W delay product (BDP)	≈ 200 pkts (Exp1) ≈ 120 pkts (Exp2)
Bottleneck Buffer Size (B)	DropTail: 1.5× BDP RED/RIO-PS: 2.5×BDP
Q. Parameters	$(min_{th}, max_{th}, P_{max}, w_q)$
RED	(0.15B, 0.5B, 1/10, 1/512)
RIO-PS short	(0.15B, 0.35B, 1/20, 1/512)
RIO-PS long	(0.15B, 0.5B, 1/10, 1/512)
RED & RIO-PS	ecn_on, wait_on, gentle_on
Edge Router	$SLR = 3, T_u = 1\ sec, T_c = 10\ sec$
Foreground Traffic	
(Src, Dest)	(Server Pool, Client Pool)
Long Connection Size	1000 packets
Short Connection Size	10 packets

TABLE III

NETWORK CONFIGURATION

Simulation Duration

- Experiments run 4000 seconds with a 2000 second warm-up period.
- Why??

Figure 6a. Relative Response Time [RIO = 3 sec.]

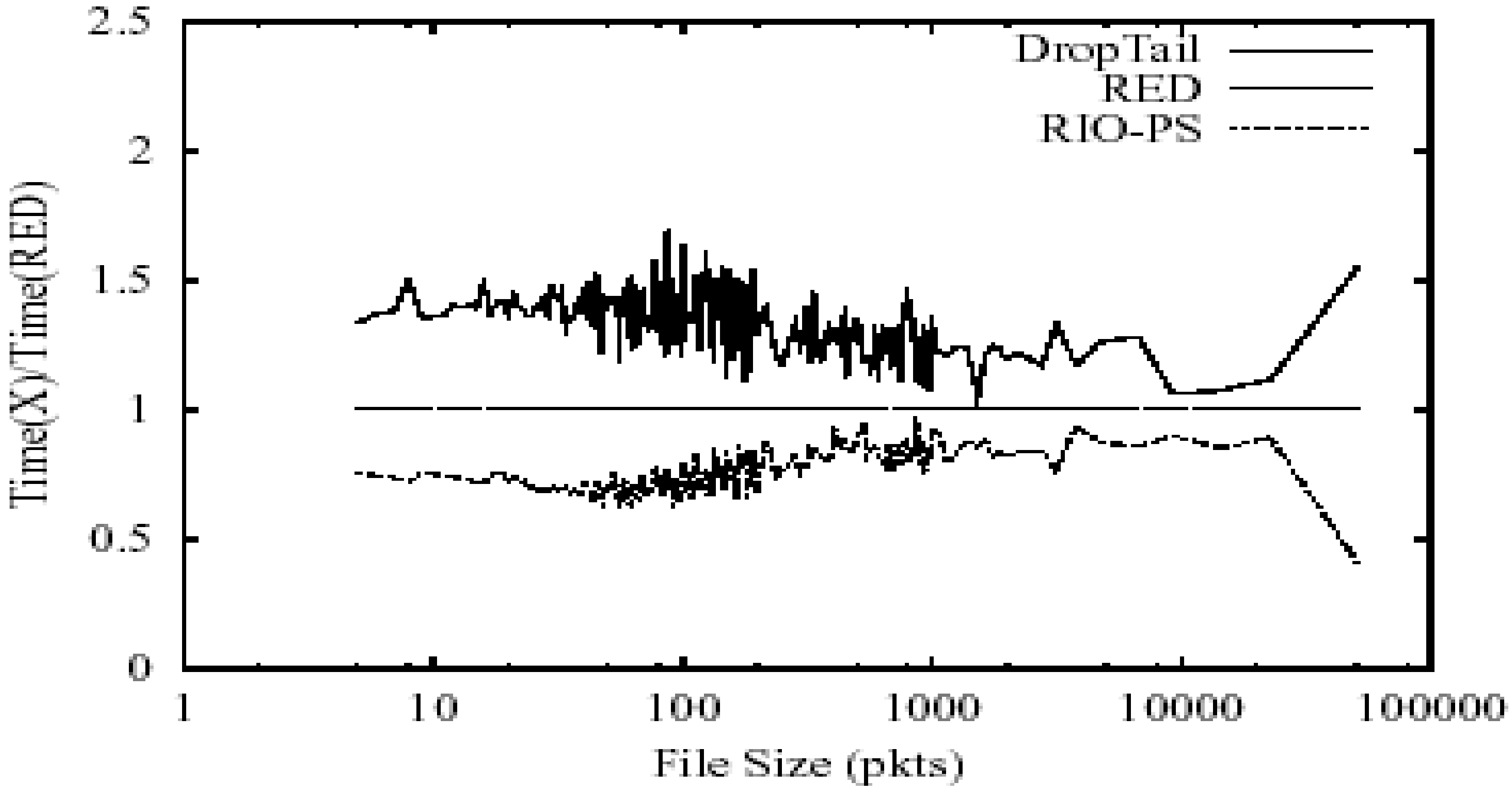


Figure 6b. Relative Response Time [RIO = 1 sec.]

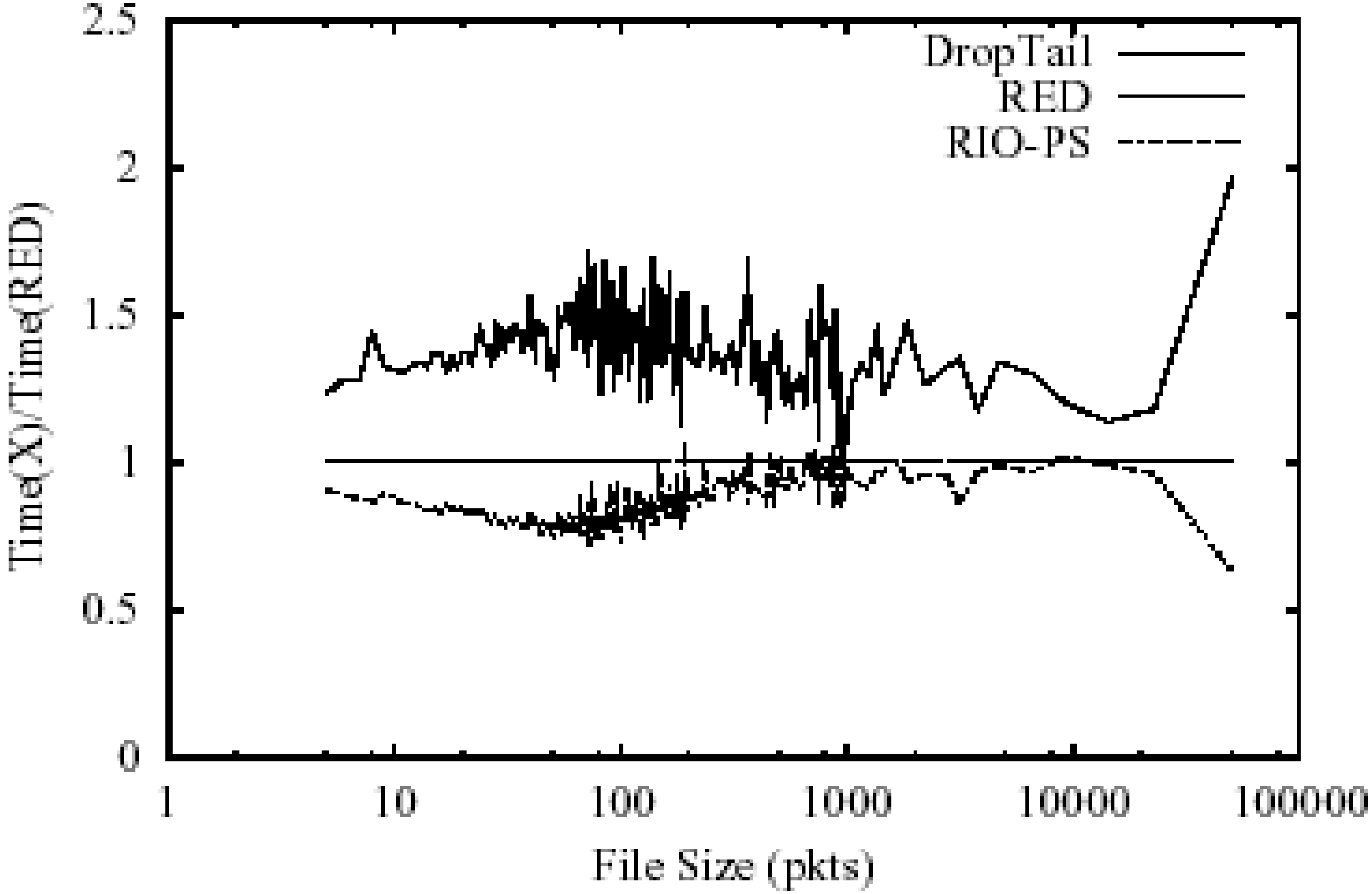


Figure 7a. Instantaneous Queue Size

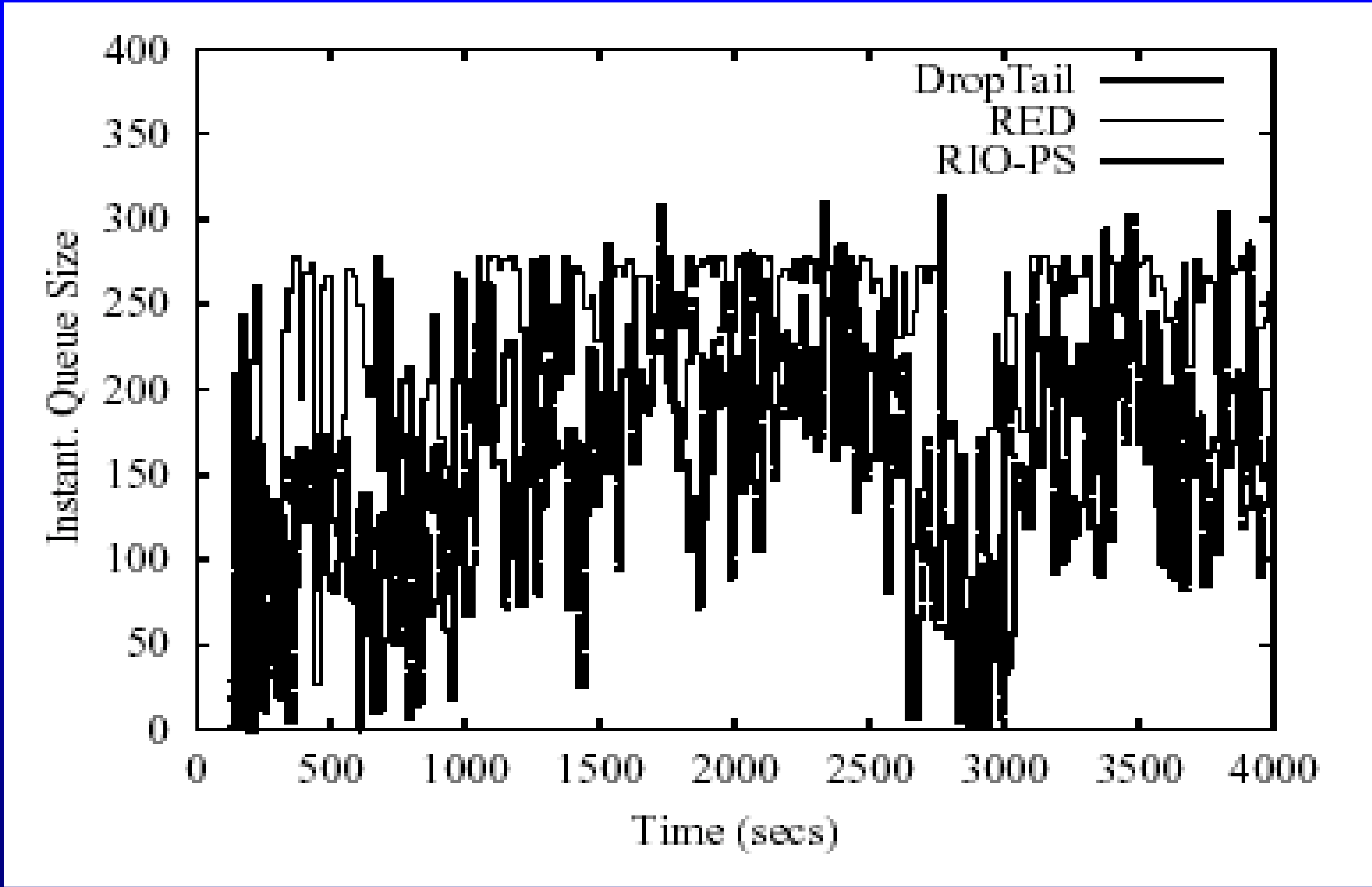
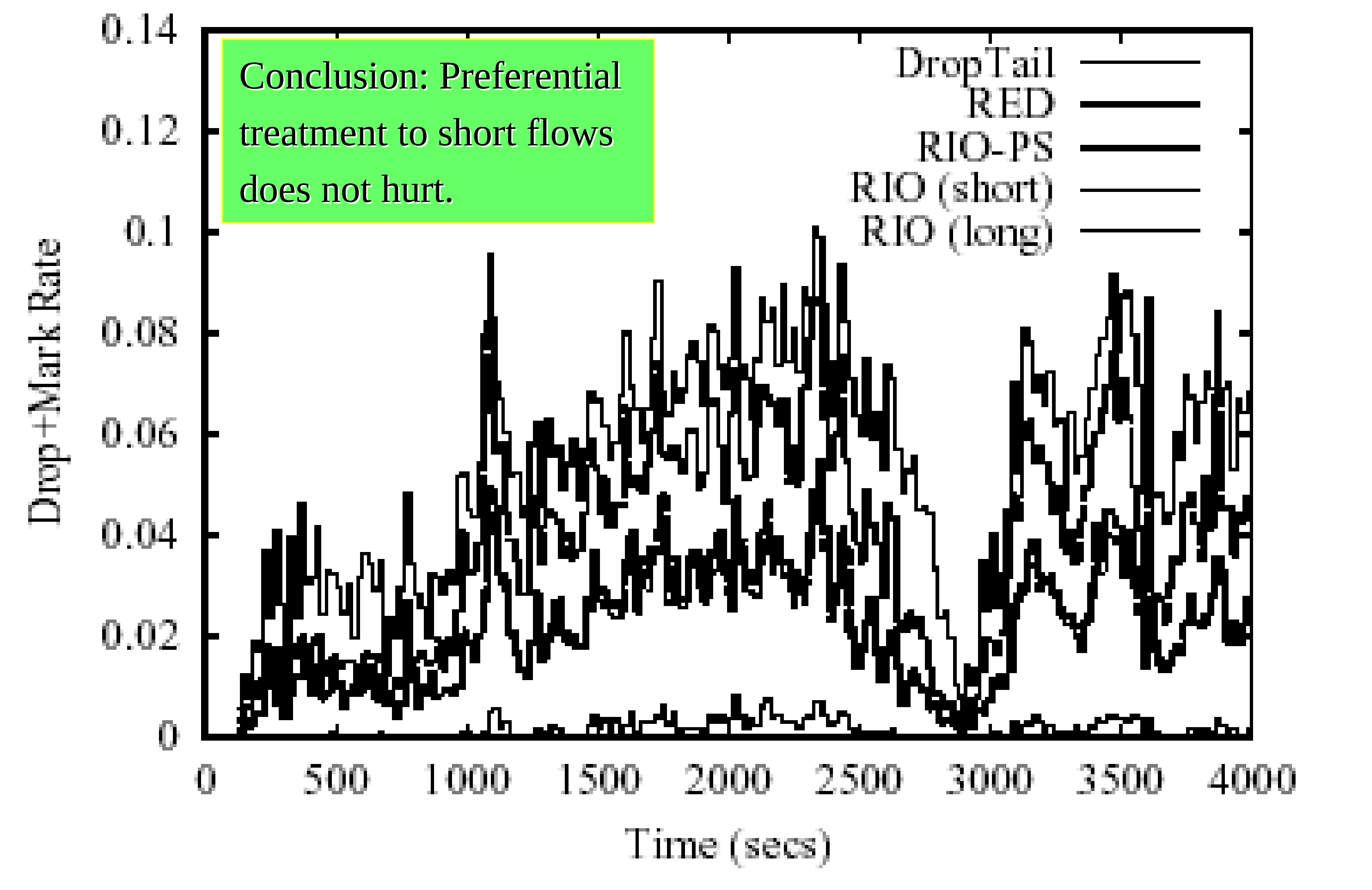


Figure 7b. Instantaneous Drop/Mark Rate



Foreground Traffic Study

- Periodically injected 10 short flows (every 25 seconds) and 10 long flows (every 125 seconds) as foreground TCP connections and recorded the response time for i^{th} connection.

Figure 8a. Jain's Fairness – Short Connections

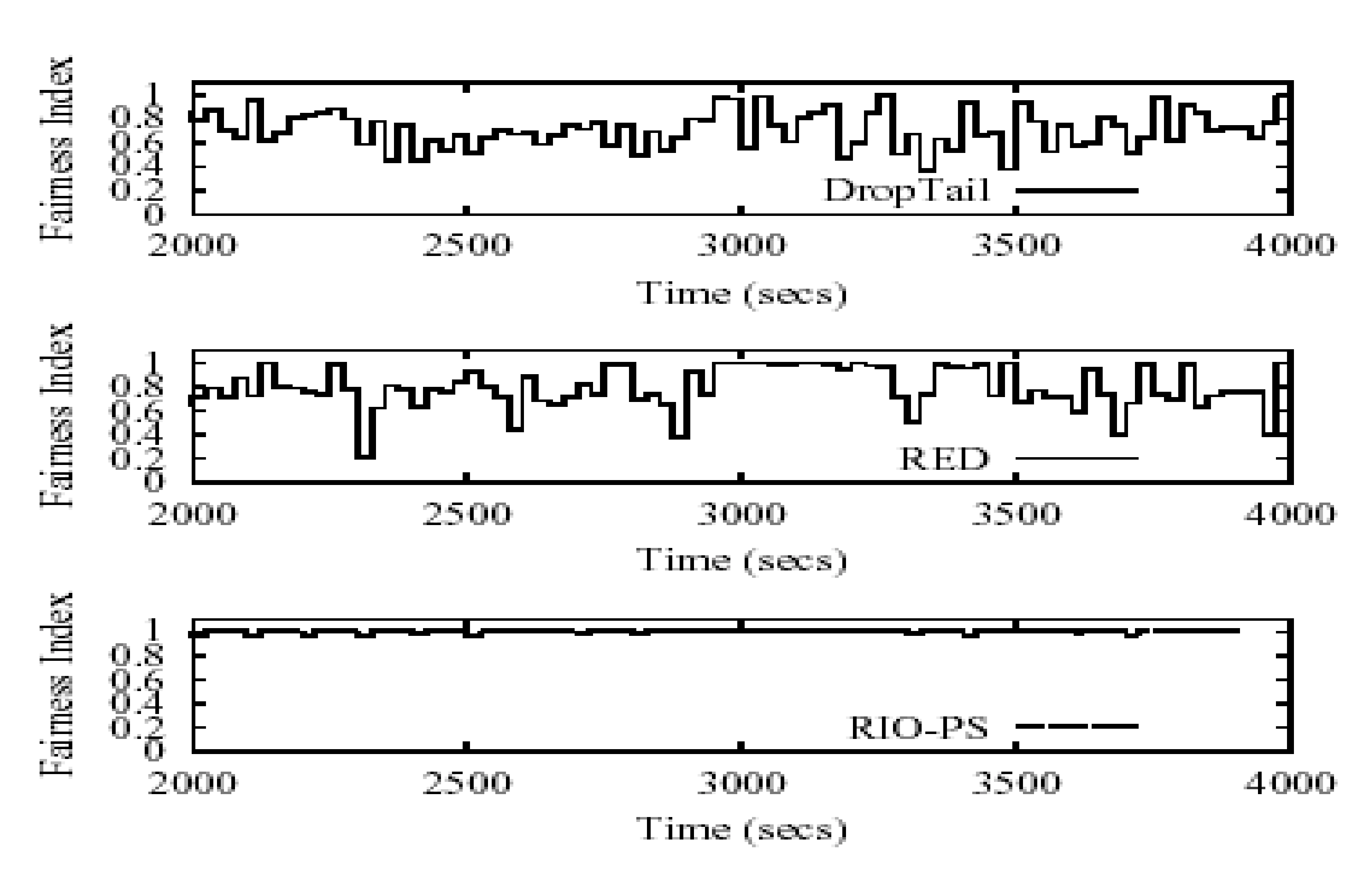


Figure 8b. Jain's Fairness – Long Connections

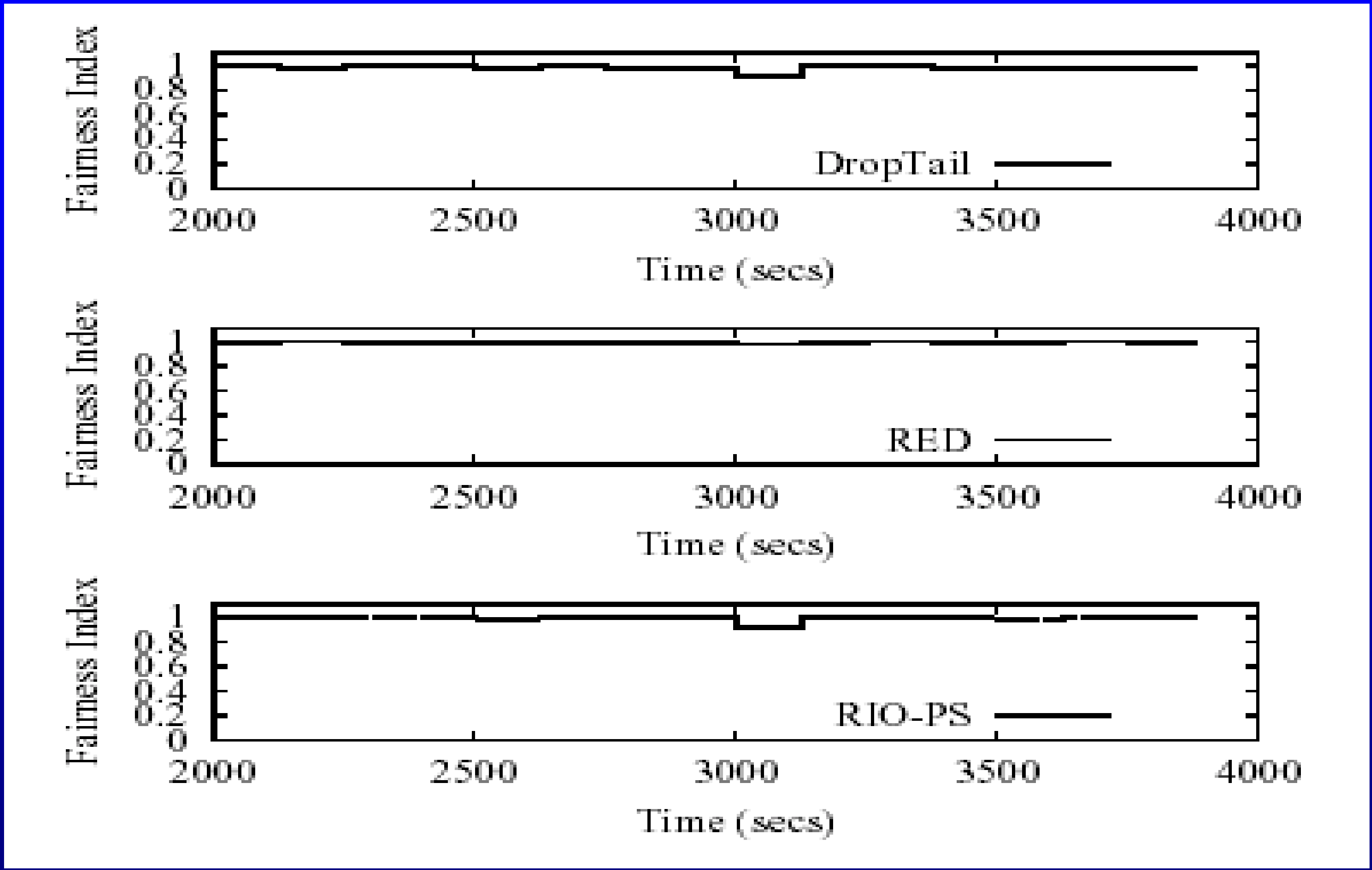


Figure 9a. Transmission Time – Short Connections

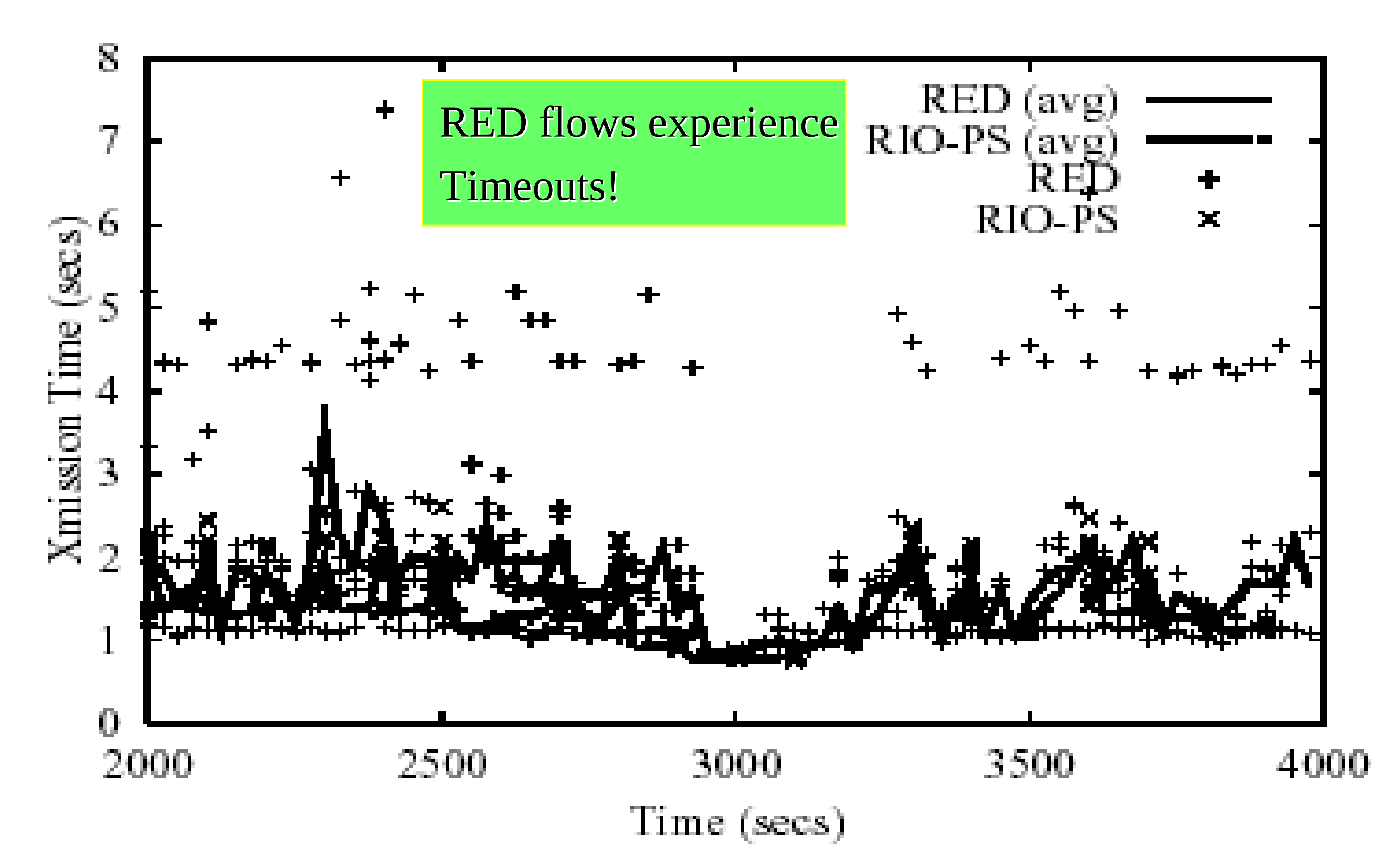


Figure 9b. Transmission Time – Long Connections

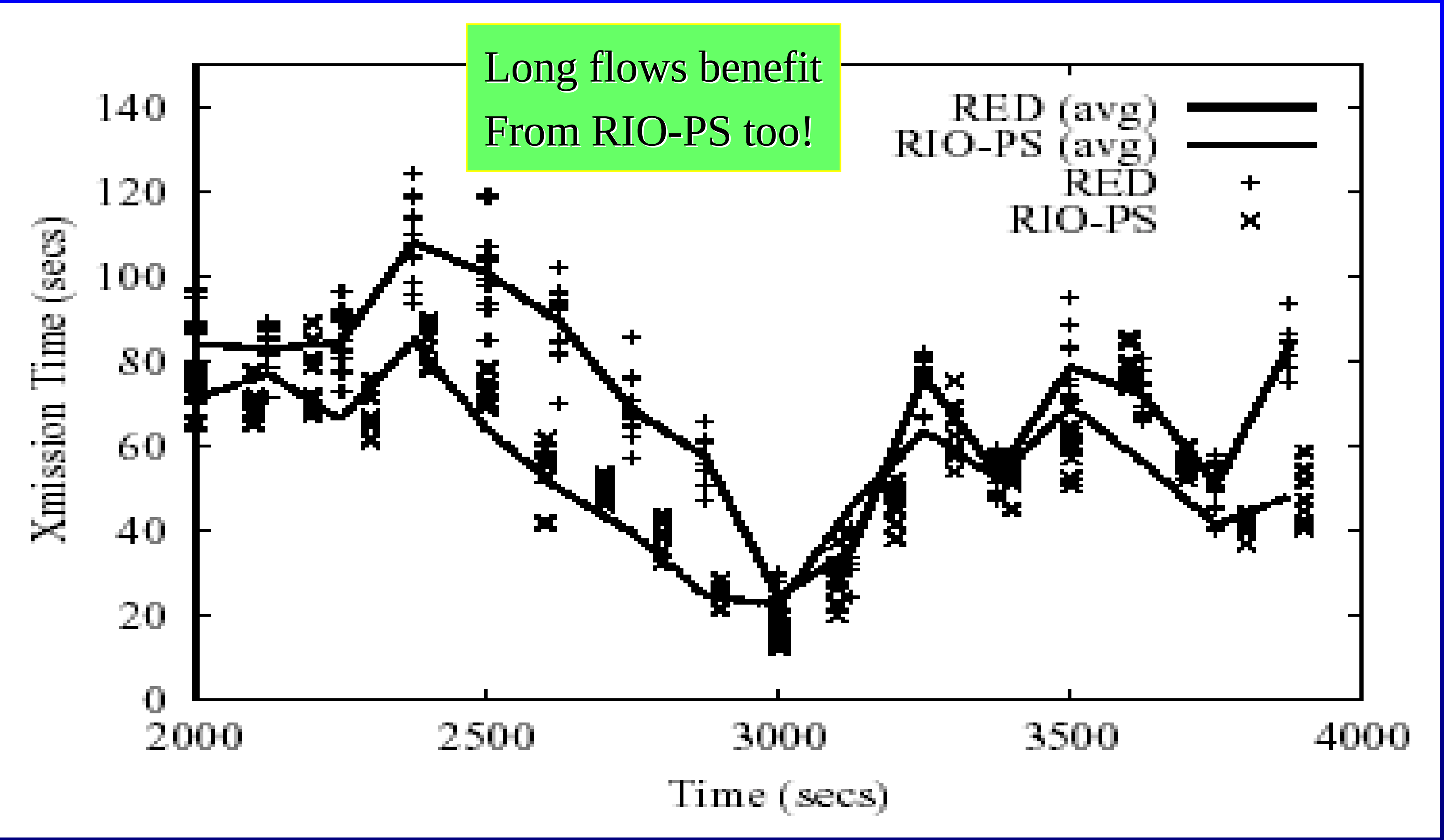


Table IV
Network Goodput over the Last 2000 secs.

Scheme	DropTail	RED	RIO-PS
Exp1 (ITO=3sec)	4207841	4264890	4255711
Exp1 (ITO=1sec)	4234309	4254291	4244158
Exp2 (ITO=3sec)	4718311	4730029	4723774

Discussion

- Only did one-way traffic. Authors claim two-way would be even better for RIO-PS.
- Argument: Others have shown that edge routers do not significantly impact performance.

Conclusions

- Proposed architecture with edge routers classifying flows and core routers implementing RIO-PS.
- This scheme shown to improve response time and fairness for short flows.
- The performance of long flows is also enhanced.
- Overall goodput is improved {a weak claim}.
- Authors call their approach "size-aware" traffic management.